International Journal of Peptide Research and Therapeutics (2021) 27:309–316
https://doi.org/10.1007/s10989-020-10087-7

Using Chou's Five-steps Rule to Classify and Predict Glutathione S-transferases with Different Machine Learning Algorithms and Pseudo Amino Acid Composition

Hassan Mohabatkar[1] · Samira Ebrahimi[1] · Mohammad Moradi[1]

Accepted: 8 June 2020 / Published online: 13 June 2020

Abstract
The Glutathione S-Transferases (GSTs) are detoxification enzymes which exist in variety of living organisms such as bacteria, fungi, plants and animals. These multifunctional enzymes play important roles in the biosynthesis of steroids, prostaglandins, apoptosis regulation, and stress signaling. In this study, we designed a method to independently predict the structures of animal, fungal and plant GSTs using Chou's pseudo-amino acid composition concept. Support vector machine (SVM), Random Forests (RF), Covariance Discrimination (CD) and Optimized Evidence-Theoretic K-nearest Neighbor (OET-KNN) were used as powerful machine learnings algorithms. Based on our results, Random Forests demonstrated the best prediction for animal GSTs with 0.9339 accuracy and SVM showed the best results for fungal and plant GSTs with 0.8982 and 0.9655 accuracy, respectively. Our study provided an effective prediction for GSTs based on the concept of PseAAC and four different machine learning algorithms.

Keywords Glutathione S-Transferases · Chou's PseAAC · Machine learning

Introduction

Glutathione S-Transferases (GSTs) are known as detoxification enzymes, which detoxify reduced form of glutathione toxenobiotic substrates (Dasari et al. 2018). However, these proteins have other cellular defense functions; such as a metabolic roles in the biosynthesis of steroids, prostaglandins (Landi 2000), aromatic amino acid degradation (Kilty et al. 1998), apoptosis regulation, and stress signaling (Tew and Ronai 1999).

GSTs are multifunctional, heterogeneous superfamily of proteins that catalyze the conjugation of GSH by a sulfhydryl group (–SH) to electrophilic centers and can catalyze different reactions with their distinct substrates. GST are divided into eight subgroups based on their amino acid sequences, isoelectric points and immunological properties; including alpha, mu, phi, theta, kappa, zeta, sigma, and omega. Most of GSTs contain homodimers with 23–30 kDa subunits. Each GST has two typical domains; one conserved N-terminal, a glutathione-specific binding domain (G-site), and a variable C-terminal domain which binds to the hydrophobic substrate (C-site) (Schultz and Sylvester 2001). GSTs are available in a variety of livings organisms, from bacteria to homo sapiens (Di Matteo et al. 2019). Plant GSTs are six classes of a multifunctional heterogeneous superfamily that perform a wide range of non-enzymatic and pivotal catalytic functions (Sylvestre-Gonon et al. 2019). The earliest evidence for GSTs in mammalians came from the discovery of GSTs in rat liver, early 1960s. They are classified into three different groups; including membrane-bound microsomal, mitochondrial and cytoplasmic. Mammalian GSTs have various activities; from detoxification of reactive electrophiles to cell signaling (Landi 2000).

Fungal GSTs have been discovered in a number of fungal species. Some GSTs could play critical roles in fungal necrotrophs protection against reactive oxygen species and plant-derived toxic metabolites that aggregate on the host-pathogen interface during infection. In addition, GSTs could bind wood-derived molecules and are able to participate in the intracellular transport and further omission of these compounds, which could be toxic for the cells (Kato et al. 2004).

✉ Hassan Mohabatkar
h.mohabatkar@ast.ui.ac.ir

[1] Department of Biotechnology, Faculty of Biological Science and Technology, University of Isfahan, Isfahan, Iran

However, the biological role and importance of GTSs are not yet completely understood.

Since the implementation of the human genome project in 2000, and benefited from the next-generation sequencing technology, a huge number of bioinformatics data have been accumulated, which is far beyond the capacity of classical molecular analysis (Roberts et al. 2006). Bioinformatics technology can provide a practical and time-saving approach to answer critical research questions and find a correlation between data. Formulating the sequence of biological samples; such as protein or RNA, effectively with a separate model that can considerably keep its sequence order information, is one of the most difficult challenges in computational biology.

Nowadays, different aspects of proteins are predicted by various approaches. Some of these methods are based on amino acid composition (Lee et al. 2006); sequence (Xia et al. 2010), and template (Chen and Kihara 2011; Haghighi et al. 2019). Pseudo amino acid composition is a typical discrete form that has been widely used for protein function prediction, which presents information based on the protein sequence. The pseudo amino acid (PseAA) composition can provide a lot more information about a protein sequence compared with the conventional amino acid (AA) composition (Chou 2019; Chou et al. 2019; Tian et al. 2019; Yu et al. 2017).

Machine learning is a method for data analysis that automates analytical model making (Chen et al. 2020). This branch of technology allows the systems to learn from data, recognize patterns and make decisions with minimal human intervention (Galetsi et al. 2020). Machine learning algorithms enable the system to learn from data and make a prediction based on those information (Chakrabarti et al. 2006). Support-vector machine (SVM) is a novel kernel-based machine learning which is useful for biological researches (Schölkopf et al. 2002). Random Forest (RF) is another learning method for classification which operates by building a multitude of decision tree at training time and outputting the classification or prediction (Kam 1995). Covariance Discrimination (CD) and optimized evidence-theoretic K-nearest neighbor (OET-KNN) are two other strong learning methods which are used in the classification of biological data (Shen and Chou 2005). In the present investigation, we aimed to predict and classify GSTs by the concept of Chou's PseAAC. SVM, RF, CD, and OET-KNN that were used as classifiers algorithms to predict GSTs. It has been proposed that five-steps rule of Chou is a notable achievement in the field of molecular biology (Chou 2020). This method has been used in many different studies such as predicting remote homology proteins through merging grey incidence analysis and domain similarity investigation (Lin et al. 2020).

In order to make an appropriate biological prediction, the "Chou's 5-steps rule" (Chou 2011) could be used which includes (to see the example, please refer to reference (Chou 2019) of this study); first, selection or building up an appropriate benchmark dataset for testing and training of the prediction. The second step is providing the samples with efficient formulation which could mirror their inherent interaction with the objective of prediction. The third level of Chou's five step is presenting a strong algorithm for conducting the prediction process, and forth one is appropriate performance of the cross-validation assays to measure the prediction's accuracy. The final step in the mentioned method is establishing an easy to use web-based tool for the predictor with public access. Articles written based on this usually provide an easily understood logic of development, high transparency in work, the results could be easily repeated by the other researchers, they could prompt the rest of approaches for analyzing the sequences, and provide a comfortable approach for a large group of experimental researchers.

Methodology

Datasets Preparation for Analyzing GSTs Pseudo Amino Acids Composition

Three primary datasets consisting 671 fungal protein sequences, 586 animal protein sequences and 703 plant protein sequences of GSTs, as the positive dataset and 638, 829 and 681 sequences of non-GSTs proteins, respectively as the negative fungal, animal and plant datasets were fetched. These sequences were obtained from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/protein). To increase the accuracy of prediction, similar and fragmental sequences were not chosen. To avoid intolerance, the sequences with more than 90% similarity were excluded from our datasets, using CD-Hit tool (Li and Godzik 2006). Final positive fungal, animal and plant positive datasets were including 584, 513 and 541 sequences, respectively. Also, negative dataset sequences for fungal, animal and plant non-GSTs proteins were decreased to 523, 576 and 501, respectively.

Generating Chou's PseAAC for GSTs Protein

In order to operate datasets for data mining approaches, proteins were presented numerically using the PseAAC web server (Shen and Chou 2008). Kou-Chen Chou introduced the concept of PseAAC in 2001 which represents the protein attributes according to the sequence-order information, enabling to provide a comprehensive combination with other properties to perform a reliable classification (Chou and Cai

2003). First 20 numbers reflect the amino acid composition and additional numbers demonstrate sequence-order information. The PseAAC web server has the following six amino acid characters: hydrophilicity, hydrophobicity, mass, pK2, pK1 and pI. Total 63 combinations of six characters were used. For the present study, type 1 PseAAC, which is also called the parallel correlation type was used. In order to provide the most reliable results, the weight factor was set on different figures from 0.1 to 1.0 and ƛ parameter was set on the valid range of numbers from 1 to 10.

Classification of GSTs by Machine Learning Algorithms

Nowadays, a practical way to explore patterns in biological data is to employ machine learning algorithms. Machine learning is a sub-branch of computer science analyzing assessment of algorithms that can be used for prediction and classification of data based on the models learned from sample input (Li et al. 2018). In machine learning algorithms, a training data is needed in order to predict without being programmed to perform the task. There are two main categories in the field of machine learning which are supervised and unsupervised learning. Supervised learning utilizes a dataset whose features are characterized. In unsupervised learning, no labels are assigned to the algorithm and the goal of this task is to explore existing data and identify the similarities. In this study, the data were assigned by known attributes; therefore, supervised learning was applied (Snoek et al. 2012).

There are many algorithms which have satisfying results for biological data. In the present study, Rapid Miner software (version 7.2) was used for validation and classification of data (Snoek et al. 2012). The following algorithms incorporated in the software were selected for GST classification:

Support Vector Machine (SVM) Model

SVMs are machine learning algorithms that analyze data used for regression and classification analysis. The goal of SVM algorithm is to find the most appropriate classification function to differentiate between members of two classes in a training dataset (Suykens and Vandewalle 1999).

Random Forest (RF) Model

RFs were initially introduced by Breiman in 2001 (Breiman 2001). This algorithm adds extra randomness layers to the bagging. RF is an ensemble learning method for regression and classification. RFs operate tasks that construct a multitude of decision trees at training time and output classification or prediction of the individual trees. In comparison with individual classification trees, RFs can highly improve the accuracy of the predictions (Pal 2005).

Covariance Discrimination (CD) Algorithm

Covariance matrix uses a standardized scale of − 1 to + 1 to achieve the strength of a relationship between two variables. Covariance is a measure of how the changes in one variable are connected with changes in the second one. Specifically, covariance calculates the degree to which two variables are linearly associated (Zou et al. 2020).

OET-KNN Algorithm

The optimized evidence-theoretic K-nearest neighbor (OET-KNN) which is a non-parametric learning algorithm for data classification (Ghosh et al. 2020), was used to separate the data related to GST proteins into different classes, in order to predict the classification and regression of a new sample dataset.

Results

In this study, parameters of sensitivity (SEN), specificity (SPC), accuracy (ACC) and Matthew's correlation coefficient (MCC) were used for the evaluation of classifiers' performances, computed based on the following equations:

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$SPC = TN/(TN + FP) \quad (2)$$

$$SEN = TP/(TP + FN) \quad (3)$$

$$MCC = [(TP * TN) - (FP * FN) / \sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (4)$$

TP, TN, FP, and FN are the numbers of true positives (GSTs predicted as GSTs), true negatives (non- GSTs predicted as non- GSTs), false positives (non- GSTs predicted as GSTs) and false negatives (GSTs predicted as non- GSTs), respectively. The overall performance of the method was obtained by averaging the above parameters for all classifiers (SEN, SPEC, ACC, and MCC). Accuracy (ACC) is the number of those that have been classified correctly. Specificity (SPC), known as true negative rate, calculates the ability to predict negative data. Therefore, results with high

SPC are reliable. High SEN of a classifier indicates negative predicted results that are reliable. MCC or Matthew's correlation coefficient is used for the evaluation of binary classification and is a number between −1 and +1, where +1 indicates an ideal classification, 0 represents random classification and −1 means complete dissent between prediction and observation. When values of these four parameters are more than 70–80%, the prediction is believed to be highly accurate. Performance values of classifiers were optimized and best results for each classifier were reported. Along with them, ƛ and *W* parameters of machine learning were taken into consideration for classification's efficacy of each dataset.

Every algorithm was trained with the dataset individually. This validation technique divides the dataset into training and testing sets. A variety of rounds of validation are applied and the results are averaged. Using cross-validation, a more accurate evaluation of the performance of model prediction was achieved. Algorithms with the best results were selected for further validation with the independent test dataset.

Optimized results provided by various machine learning algorithm for animal, plant and fungal sources are provided in Tables 1, 2 and 3. The highest values for each parameter is indicated in bold. All four machine learning algorithms had a good total accuracy higher than 0.62, certifying the provided results. According to the results provided at this section, RFs demonstrated the best prediction for animal GSTs with 0.9339 accuracy and SVM showed the best results for fungal and plant GSTs with 0.8982 and 0.9655 accuracies, respectively.

Machine Learning Based Classification for Animal GSTs

Random forest algorithm presented the highest values for the classification of animal cell derived GST enzyme including; ACC, MCC, and SEN. The highest SPC was presented by SVM algorithm, along with second-high ACC, MCC, and SEN. The ƛ was variable among different models, with a value of eight for SVM, and the lowest value of one for the CD algorithm. From the perspective of *W* parameter, the CD algorithm had the highest value, equal to 0.5, and all the others showed values equal to 0.1.

Among the used algorithms, CD showed the lowest values for ACC (0.79), MCC (0.60), and SPC (0.74). The OET-KNN algorithm demonstrated the lowest SEN of all the algorithms with a value of 0.81. Table 1 provides the detail of all the results provided by each algorithm, which could further interest the readers.

Table 1 The performance of various machine learning algorithms for animal GSTs classification

Animal GSTs						
Machine Learning Algorithms	ƛ	*W*	ACC	MCC	SEN	SPC
SVM	**8**	0.1	0.9275	0.8553	0.8928	**0.9583**
Covariance Discrimination	1	**0.5**	0.7971	0.6005	0.8519	0.7483
Random Forests	7	0.1	**0.9339**	**0.8693**	**0.961**	0.9097
OET-KNN	3	0.1	0.8898	0.7835	0.8168	0.9549

Table 2 The performance of various machine learning algorithms for plant GSTs prediction

Plant GSTs						
Machine Learning Algorithms	ƛ	*W*	ACC	MCC	SEN	SPC
SVM	4	0.1	**0.9655**	**0.9308**	0.9614	**0.9703**
Covariance Discrimination	1	0.1	0.8004	0.637	**0.9778**	0.6088
Random Forests	**8**	0.1	0.9395	0.879	0.9353	0.9441
OET-KNN	2	0.1	0.8762	0.7671	0.7911	0.9681

Table 3 The performance of various machine learning algorithms for fungal GSTs prediction

Fungal GSTs						
Machine Learning Algorithms	ƛ	*W*	ACC	MCC	SEN	SPC
SVM	**9**	0.1	**0.8982**	**0.7983**	0.8716	**0.9278**
Covariance Discrimination	4	0.1	0.6243	0.3405	**0.9949**	0.2129
Random Forests	7	0.1	0.8838	0.7687	0.9298	0.8327
OET-KNN	2	0.1	0.7703	0.5715	0.6404	0.9144

Machine Learning Based Classification for Plant GSTs

For the classification of the GST enzyme of the plants, SVM presented the highest values for ACC (0.96), MCC (0.93), and SPC (0.97). After SVM in the second rank, the Random forest showed the highest ACC, and MCC values. The highest sensitivity was for the CD model (0.9778) and the OET-KNN indicated the second-high specificity for plant GST classification.

The *W* value was the same for all the algorithms, equal to 0.1. From the ϰ parameter point of view, the Random forest had the highest value (eight) and the CD showed the lowest one with a value equal to one. The lowest ACC, MCC, and SPC value for classification of plant GST was presented by CD algorithm, and similar to the animal GST classification, the KNN showed the lowest SEN value. The detailed results ofant derived GSTs are provided in Table 2.

Machine Learning Based Classification for Fungal GST

The SVM model provided the highest values in accuracy (0.89), specificity (0.92), and Matthew's correlation coefficient (0.79) for classification of fungal GST enzyme, but the highest sensitivity was provided by the CD model (0.99). The second-high ACC (0.88), and MCC (0.76) and SEN (0.92) values were provided by the RF algorithm. Similar to plant classification, the second-high SPC (0.91) was provided by the OET-KNN model. The *w* parameter was the same among all the chosen models, and it was equal to 0.1. The ϰ parameter had different values in each model, having the highest (nine) for the SVM, and the lowest one was presented for OET-KEN model (two). Among the used algorithms, the lowest value for ACC (0.62), MCC (0.34), and SPC (0.21) belonged to CD model, and the lowest sensitivity was demonstrated by OET-KNN algorithm (0.64). The detailed results for the performed prediction of the fungal GST are provided in Table 3.

Discussion

Application of data mining methods for in silico studies has been evaluated by earlier investigations and has proven to possess a high potential for resolving the issues of old fashioned methods for categorizing and classification of biological data such as protein function domain inspection, function motif investigation, and protein function inference (Raza 2012). The importance of machine learning algorithms for data classification is beyond any question and many different area of science, including molecular biology have been using these methods for classification of the large datasets for various molecules such as enzymes (Gupta et al. 2019).

GSTs are detoxification enzyme with a wide range of known activities, available in various of living organisms such as fungal, animal and bacteria. Because of the importance of GSTs, it is of critical importance to predict and classify its different subtypes from different organisms; therefore, information about the efficacy of different algorithms of machine learning will be useful for further investigation about this enzyme and might be used for designing a server for such data classification. In the previous reports, machine learning based methods have been used for enzymes classification and clustering (Yadav and Tiwari 2015). At the current investigation, we applied machine learning models including OET-KNN, Random Forests, SVM, and Covariance Discrimination for the PseAAC data analyzing and interpretation.

Due to the large amount of data on protein sequences, investigating the function of these sequences is necessary. Most of the current methods have been developed to predict the function of proteins are based on alignment and similarity of the sequences. These methods are error-prone in the absence of significant similarity and variation in the size of queried and the target protein and most of the existing machine-learning algorithms can only handle vectors as elaborated in a comprehensive review (Chou 2015). One of the most reliable approaches for protein classification is Chou's PseAAC (Du et al. 2014). To prevent the complete loss of sequence-pattern information for proteins, the pseudo amino acid composition (Chou 2001) or PseAAC (Chou 2005) has been developed. Upon introduction of Chou's PseAAC, this method has been used in various fields of computational proteomics (please refer provided examples in reports (Behbahani et al. 2019; Chou 2017; Dehzangi et al. 2015)).

Due to a large application demand, four strong and free access tools as BPseAAC^, BPseAAC-Builder^, Bpropy^, and BPseAAC-General^, are provided by the developer (Cao et al. 2013; Du et al. 2014; Du et al. 2012; Shen and Chou 2008). The initial three programs are for creating various modes of Chou's exclusive PseAAC (Chou 2009), such as "Functional Domain" mode (see Eqs. 9–10 of (Chou 2011)), "Gene Ontology" mode (see Eqs. 11–12 of (Chou 2011)), and "Sequential Evolution" or "PSSM" mode (see Eqs. 13–14 of (Chou 2011)), though the BPseAAC-General^ is useful for Chou's general PseAAC (Chou 2011).

This method has been used in many protein prediction studies such as prediction of GABA receptor (Mohabatkar et al. 2011), protein cellular attributes (Chou 2001), cyclin proteins (Mohabatkar 2010), risk type of human papillomaviruses (Esmaeili et al. 2010), outer membrane proteins (Lin 2008) and secondary structure content (Chen et al. 2009). Due to the successful achievements of using PseAAC for protein/

peptide sequence processing, PseKNC (Pseudo K-tuple Nucleotide Composition) (Chen et al. 2014) has been advanced to create plenty of feature vectors for DNA/RNA sequences (Chen et al. 2015, Liu et al. 2018) which have proven to be very practical. These methods could be applied for providing any desired feature vectors for protein/peptide and DNA/RNA sequences in adjustment with the requirements of studies. Particularly, in 2015 a very powerful web-server called 'Pse-in-One' (Liu et al. 2015) and its updated version 'Pse-in-One2.0' (Liu et al. 2017) have been established that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

Although GSTs have a crucial role in survival from cancer and other diseases (Allocati et al. 2018), there is few reports on approaches for predicting the GSTs proteins. One of the most highlighted researches in this field was the study by Nitish Kumar Mishra et al. who initially used a SVM based method has been developed using amino acid and dipeptide composition and achieved the maximum accuracy of 91.59% and 95.79%, respectively. In their study, they designed a SVM-based method using tripeptide composition and achieved maximum accuracy 97.66%, which is showed to be higher than the accuracy achieved by HMM based searching (96.26%). Based on the results of their study, they developed a web-server named GSTPred (http://www.imtech.res.in/raghava/gstpred/) (Mishra et al. 2007). The results provided by the current machine learning study also showed that SVM could provide high scores of ACC, SPC and MCC for the classification of fungal and plant GSTs, but the RF algorithm showed highest scores of ACC, MCC, and SEN for the classification of animal GSTs. It is also necessary to mention that Covariance Discrimination model could provide a high sensitivity for classification of fungal and plant GST.

There is no pre-existing server available for predicting different types of GSTs. Here our results demonstrate that using the concept of Chou's PseAAC and different machine learning algorithms is a successful method in predicting different types of GSTs. Next step of this study could be designing a web server to predict animal, fungal and bacterial GSTs proteins. In the present investigation, a mixture of amino acid composition and four different machine learning algorithms for distinguishing fungal, animal and bacterial GSTs were applied. SVM method shows the highest accuracy for plant and fungal GSTs (96.55% and 89.82%, respectively). Also, RF shows 93.39% accuracy for the animal GSTs.

As Chou et al. stated in their report (Chen et al. 2016; Liu et al. 2016), user-friendly and free access web programs prepare guidance for future bioinformatics investigations (examples provided at (Chouet al. 2019; Xiao et al. 2019)). Indeed, they have greatly increased the impact of in silico methods on medical science (Chou 2015, 2019), which has provided a significant breakthrough in medical sciences (Chou 2017). Using these valuable insights, we wish to use the provided results of the current study to continue this investigation and design an adjustable program for such a data classification by application of the mentioned algorithms.

Conclusion

In conclusion, due to the importance of GSTs, prediction of fungal, plant and animal GSTs was performed using the concept of Chou's PseAAC. Results indicated high accuracy using different machine learning algorithms for classifying this enzyme by their sequences and shed light on the variation among different models. Such predictions and classifications are critically important before choosing the best source of the enzyme for specific applications.

Acknowledgements Support of this study by the University of Isfahan is acknowledged.

Compliance with Ethical Standards

Conflict of interest There is no conflict to declare.

Informed Consent There was no human participant and consent was not required.

Research involving Human and/or Animals Participants No human or animal was participated in this study.

References

Allocati N, Masulli M, Di Ilio C, Federici L (2018) Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases. Oncogenesis 7:1–15

Behbahani M, Nosrati M, Moradi M, Mohabatkar H (2019) Using Chou's general pseudo amino acid composition to classify lac-cases from bacterial and fungal sources via Chou's five-step rule. Appl Biochem Biotechnol 190:1035–1048

Breiman L (2001) Random forests. Machine Learn 45:5–32

Cao D-S, Xu Q-S, Liang Y-Z (2013) propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962

Chakrabarti S, Ester M, Fayyad U, Gehrke J, Han J, Morishita S et al (2006) Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee

Chen H, Kihara D (2011) Effect of using suboptimal alignments in template-based protein structure prediction. Proteins Struct Funct Bioinf 79:315–334

Chen C, Chen L, Zou X, Cai P (2009) Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett 16:27–31

Chen W, Lei T-Y, Jin D-C, Lin H, Chou K-C (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. Anal Biochem 456:53–60

Chen W, Lin H, Chou K-C (2015) Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst 11:2620–2634

Chen W, Tang H, Ye J, Lin H, Chou K-C (2016) iRNA-PseU: identifying RNA pseudouridine sites. Mol Ther Nucleic Acids 5:e332
Chen Y, Banerjee D, Mukhopadhyay A, Petzold CJ (2020) Systems and synthetic biology tools for advanced bioproduction hosts. Curr Opin Biotechnol 64:101–109
Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinf 43:246–255
Chou K-C (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou K-C (2009) Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteomics 6:262–274
Chou K-C (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247
Chou K-C (2015) Impacts of bioinformatics to medicinal chemistry. Med Chem 11:218–234
Chou K-C (2017) An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr Topics Med Chem 17:2337–2358
Chou K-C (2019) Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. Curr Med Chem 26:4918–4943
Chou K-C (2020) Proposing 5-steps rule is a notable milestone for studying molecular biology. Nat Sci 12:74
Chou KC, Cai YD (2003) Predicting protein quaternary structure by pseudo amino acid composition. Proteins Struct Funct Bioinf 53:282–289
Chou K-C, Cheng X, Xiao X (2019) pLoc_bal-mEuk: predict subcellular localization of eukaryotic proteins by general PseAAC and quasi-balancing training dataset. Med Chem 15:472–485
Dasari S, Ganjayi MS, Yellanurkonda P, Basha S, Meriga B (2018) Role of glutathione S-transferases in detoxification of a polycyclic aromatic hydrocarbon, methylcholanthrene. Chemico-Biol Interact 294:81–90
Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A (2015) Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou′ s general PseAAC. J Theor Biol 364:284–294
Di Matteo A, Federici L, Masulli M, Carletti E, Santorelli D, Cassidy J et al (2019) Structural characterization of the Xi Class glutathione transferase from the Haloalkaliphilic Archaeon Natrialba magadii. Front Microbiol 10:9
Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem 425:117–119
Du P, Gu S, Jiao Y (2014) PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci 15:3495–3506
Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. J Theor Biol 263:203–209
Galetsi P, Katsaliaki K, Kumar S (2020) Big data analytics in health sector: theoretical framework, techniques and prospects. Int J Inf Manag 50:206–216
Ghosh C, Saha S, Saha S, Ghosh N, Singha K, Banerjee A et al (2020) Machine Learning Based Supplementary Prediction System Using K Nearest Neighbour Algorithm. Available at SSRN 3517197
Gupta CLP, Bihari A, Tripathi S (2019) Protein classification using machine learning and statistical techniques: a comparative analysis. arXiv preprint arXiv:190106152
Haghighi O, Davaeifar S, Zahiri HS, Maleki H, Noghabi KA (2019) Homology Modeling and Molecular Docking Studies of Glutamate Dehydrogenase (GDH) from Cyanobacterium Synechocystis sp. PCC 6803. Int J Pept Res Ther 26:783–793
Kam HT (1995) Random decision forest. In: Proceedings of the 3rd international conference on document analysis and recognition, Montreal, Canada, 14–16 August 1995. IEEE, p 278282
Kato T, Miyakawa H, Ishibashi M (2004) Frequency and significance of anti-glutathione S-transferase autoantibody (anti-GST A1-1) in autoimmune hepatitis. J Autoimmun 22:211–216
Kilty C, Doyle S, Hassett B, Manning F (1998) Glutathione S-transferases as biomarkers of organ damage: applications of rodent and canine GST enzyme immunoassays. Chemico-Biol Interact 111:123–135
Landi S (2000) Mammalian class theta GST and differential susceptibility to carcinogens: a review. Mutat Res/Rev Mutat Res 463:247–283
Lee S, Lee B, Kim D (2006) Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins Struct Funct Bioinf 62:1107–1114
Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659
Li Y, Wu F-X, Ngom A (2018) A review on machine learning principles for multi-view biological data integration. Brief Bioinform 19:325–340
Lin H (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J Theor Biol 252:350–356
Lin W, Xiao X, Qiu W, Chou K-C (2020) Use Chou's 5-steps rule to predict remote homology proteins by merging grey incidence analysis and domain similarity analysis. Nat Sci 12:181
Liu B, Liu F, Wang X, Chen J, Fang L, Chou K-C (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res 43:W65–W71
Liu B, Fang L, Long R, Lan X, Chou K-C (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32:362–369
Liu B, Wu H, Chou K-C (2017) Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nat Sci 9:67
Liu B, Yang F, Huang D-S, Chou K-C (2018) iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics 34:33–40
Mishra NK, Kumar M, Raghava G (2007) Support vector machine based prediction of glutathione S-transferase proteins. Protein Pept Lett 14:575–580
Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett 17:1207–1214
Mohabatkar H, Beigi MM, Esmaeili A (2011) Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. J Theor Biol 281:18–23
Pal M (2005) Random forest classifier for remote sensing classification. Int J Remote Sens 26:217–222
Raza K (2012) Application of data mining in bioinformatics. arXiv preprint arXiv:12051125
Roberts E, Eargle J, Wright D, Luthey-Schulten Z (2006) MultiSeq: unifying sequence and structure data for evolutionary analysis. BMC Bioinform 7:382
Schultz IR, Sylvester SR (2001) Stereospecific toxicokinetics of bromochloro-and chlorofluoroacetate: Effect of GST-ζ depletion. Toxicol Appl Pharmcol 175:104–113
Schölkopf B, Smola AJ, Bach F (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, Cambridge
Shen H, Chou K-C (2005) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to

predict membrane protein types. Biochem Biophys Res Commun 334:288–292
Shen H-B, Chou K-C (2008) PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal Biochem 373:386–388
Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst 2:2951–2959
Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9:293–300
Sylvestre-Gonon E, Law S, Schwartz M, Robe K, Keech O, Didierjean C et al (2019) Functional, structural and biochemical features of plant serinyl-glutathione transferases. Front Plant Sci 10:608
Tew KD, Ronai ZE (1999) GST function in drug and stress response. Drug Resist Updates 2:143–147
Tian B, Wu X, Chen C, Qiu W, Ma Q, Yu B (2019) Predicting protein–protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. J Theor Biol 462:329–346
Xia J-F, Han K, Huang D-S (2010) Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. Protein Pept Lett 17:137–145
Xiao X, Cheng X, Chen G, Mao Q, Chou K-C (2019) pLoc_bal-mVirus: predict subcellular localization of multi-label virus proteins by Chou's general PseAAC and IHTS treatment to balance training dataset. Med Chem 15:496–509
Yadav SK, Tiwari AK (2015) Classification of enzymes using machine learning based approaches: a review. Machine Learn Appl 2:30–49
Yu B, Li S, Qiu W-Y, Chen C, Chen R-X, Wang L et al (2017) Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget 8:107640
Zou Q, Lin G, Jiang X, Liu X, Zeng X (2020) Sequence clustering in bioinformatics: an empirical study. Brief Bioinform 21:1–10